On Large Delays in Multi-Server Queues with Heavy Tails

Sergey Foss
Department of Actuarial Mathematics and Statistics and the Maxwell Institute for Mathematical Sciences,
Heriot-Watt University, Edinburgh EH14 4AS, Scotland and
Sobolev Institute of Mathematics, Novosibirsk 630090, Russia
email: s.foss@hw.ac.uk http://www.ma.hw.ac.uk/ams/people/pages/foss.php

Dmitry Korshunov
Sobolev Institute of Mathematics, Novosibirsk 630090, Russia
email: korshunov@math.nsc.ru http://www.math.nsc.ru/LBRT/v1/dima/dima.html

We present upper and lower bounds for the tail distribution of the stationary waiting time D in the stable $GI/GI/s$ FCFS queue. These bounds depend on the value of the traffic load ρ which is the ratio of mean service and mean interarrival times. For service times with intermediate regularly varying tail distribution the bounds are exact up to a constant, and we are able to establish a "principle of $s-k$ big jumps" in this case (here k is the integer part of ρ), which gives the most probable way for the stationary waiting time to be large .

Another corollary of the bounds obtained is to provide a new proof of necessity and sufficiency of conditions for the existence of moments of the stationary waiting time.

Key words: FCFS multi-server queue; stationary waiting time; heavy tails; large deviations; long tailed distribution; subexponential distribution; existence of moments

MSC2000 Subject Classification: Primary: 60K25, 90B22; Secondary: 60F10

OR/MS subject classification: Primary: Queues Approximations

1. Introduction and main results. We consider a *first-come-first-served* multi-server system with s identical servers. Let τ be a typical interarrival time and σ a typical service time. Independent identically distributed sequences of interarrival times $\{\tau_n\}$ with mean $a = \mathbb{E}\tau$ and service times $\{\sigma_n\}$ with mean $b = \mathbb{E}\sigma$ are assumed to be mutually independent. We also assume throughout the paper that the distribution of σ has unbounded support, i.e. $B(x) := \mathbb{P}\{\sigma \le x\} < 1$ for all x, and that the system is *stable*, i.e. $\rho := b/a < s$.

There are two equivalent ways to describe the dynamics of a multi-server system. First, we may assume that customers form a single queue in front of all servers, and that the first customer in the queue moves immediately to a server which becomes idle. Second, we may assume that customers form s individual queues (lines) – one queue for each server, service times of customers become known upon their arrival, and each arriving customer is directed to the line with a minimal total workload (we also assume that queues are numbered, and if there are more than one minimal workloads, then a customer chooses the one with the minimal number). In the rest of the paper, we mostly follow the second description of the model.

For $n = 1, 2, \ldots$, let $\mathbf{V}_n = (V_{n1}, \ldots, V_{ns})$ be the vector of residual workloads in lines $1, \ldots, s$ which are observed by the n-th customer upon its arrival into the system. The value of $D_n := \min\{V_{nj}, j \le s\}$ is the waiting time, or the delay which customer n experiences. The n-th customer joins the i_n-th line. Then

$$i_n = \min\{i : V_{ni} = D_n\}$$

and

$$V_{n+1,i} \quad = \quad \begin{cases} (V_{ni} + \sigma_n - \tau_{n+1})^+ & \text{if } i = i_n, \\ (V_{ni} - \tau_{n+1})^+ & \text{if } i \ne i_n. \end{cases}$$

Let $R(\mathbf{w}) = (R_1(\mathbf{w}), \ldots, R_s(\mathbf{w}))$ be the operator on $\mathbb{R}^s$ which orders the coordinates of $\mathbf{w} \in \mathbb{R}^s$ in the non-descending order, i.e., $R_1(\mathbf{w}) \le \cdots \le R_s(\mathbf{w})$. For $n = 1, 2, \ldots$, put $\mathbf{W}_n = R\mathbf{V}_n$. Then $D_n = W_{n1}$ and the vectors $\{\mathbf{W}_n\}$ satisfy the Kiefer–Wolfowitz [11] recursion:

$$\mathbf{W}_{n+1} = R((W_{n1} + \sigma_n - \tau_{n+1})^+, (W_{n2} - \tau_{n+1})^+, \ldots, (W_{ns} - \tau_{n+1})^+). \tag{1}$$

In a stable system, there exists a unique stationary distribution for the Kiefer-Wolfowitz vectors $\mathbf{W}_n$, and the distribution of $\mathbf{W}_n$ converges to the stationary distribution in the total variation norm, as $n \to \infty$.

In particular, the same holds for the D_n: there exists a unique distribution of the stationary waiting time (delay) D, and the distribution of D_n converges to that of D in the total variation norm.

In a single server queue ($s = 1$), the waiting times D_n satisfy the Lindley recursion [13]:

$$D_{n+1} = (D_n + \sigma_n - \tau_{n+1})^+.$$

Recall that, given $D_1 = 0$, D_{n+1} coincides in distribution with $\max(S_k, k \le n)$ where $S_0 = 0$ and $S_n = \sum_{k=1}^{n}(\sigma_k - \tau_{k+1})$, for $n \ge 1$. It is well known (see, for example, [14, 19, 1]) that the tail of stationary waiting time D is related to the service time distribution tail $\overline{B}(x) = \mathbb{P}\{\sigma > x\}$ via the equivalence

$$\mathbb{P}\{D > x\} \quad \sim \quad \frac{\rho}{1-\rho}\overline{B}_r(x) \quad \text{ as } x \to \infty, \tag{2}$$

provided the *subexponentiality* of the *residual service time distribution* B_r defined by its tail

$$\overline{B}_r(x) \quad := \quad \frac{1}{b}\int_x^\infty \overline{B}(y)\,dy, \quad x > 0$$

is guaranteed. Recall that a distribution G on $\mathbb{R}^+$ is *subexponential*, $G \in \mathcal{S}$, if $\overline{G * G}(x) \sim 2\overline{G}(x)$ as $x \to \infty$.

It is also well-known that, in a single server queue, for any $\gamma > 0$, D has a finite γth moment, $\mathbb{E}D^\gamma < \infty$ if and only if $\mathbb{E}\sigma^{\gamma+1} < \infty$, see [12]. Equivalently, $\mathbb{E}D^\gamma < \infty$ if and only if

$$\mathbb{E}\sigma_{r,1}^\gamma \quad < \quad \infty$$

where random variable $\sigma_{r,1}$ has distribution B_r.

Less is known about the stationary delay D in the multi-server queue. It is well understood that the heaviness of the stationary waiting time tail distribution depends substantially on the load ρ on the system (see, for example, the conjecture on tail equivalence by Whitt in [20]; existence results for moments in [15, 16, 17, 18]; asymptotic results for fluid queues fed by heavy-tailed on-off flows in [4, 5]). More precisely, the tail distribution depends on ρ via the value of its integer part $k = [\rho] \in \{0, 1, \ldots, s-1\}$.

For a $GI/GI/s$ system, a heuristic idea on a probable way for the large deviations to occur may be described as follows. Take $N = x\frac{k}{b-ka}$, for a very large x. Let all service times $\sigma_{n-N-s+k}, \ldots, \sigma_{n-N-1}$ be big enough, say $\sigma_{n-N-i} > x + Na$, $i = 1, \ldots, s-k$. Then the other k servers form an unstable $GI/GI/k$ queue system, because the cumulative drift of the corresponding workloads approximately equals $b - ka > 0$. In time N all workloads of these queues will exceed level x (again approximately). In this way, at time N, all s workloads become greater than x with probability which is asymptotically not less than $\overline{B}^{s-k}(x + Na) \approx \overline{B}^{s-k}\big(x\frac{b}{b-ka}\big) = \overline{B}^{s-k}\big(x\frac{\rho}{\rho-k}\big)$. We use these heuristic arguments below in Section 5 to derive a lower bound. We follow more precise calculations to obtain a better lower bound of order $\overline{B}_r^{s-k}\big(x\frac{\rho}{\rho-k}\big)$.

We recall now a few basic properties of heavy-tailed distributions and relations between them. A distribution function F is

- *long-tailed*, $F \in \mathcal{L}$, if $\overline{F}(x+1) \sim \overline{F}(x)$, as $x \to \infty$;
- *dominated varying*, $F \in \mathcal{D}$, if $\overline{F}(2x) \ge c\overline{F}(x)$, for some $c > 0$ and for all x;
- *intermediate regularly varying*, $F \in \mathcal{IRV}$, if

$$\lim_{\varepsilon \downarrow 0} \liminf_{x \to \infty} \overline{F}(x(1+\varepsilon))/\overline{F}(x) = 1;$$

- *regularly varying*, $F \in \mathcal{RV}$, if $\overline{F}(x) = l(x)x^{-\alpha}$ for $x > 0$ where $\alpha \ge 0$ is the *index* of regular variation and $l(x)$ is a *slowly varying at infinity* function, i.e. $l(cx) \sim l(x)$ as $x \to \infty$.

The following relations are known:

$$\mathcal{RV} \subset \mathcal{IRV} \subset \mathcal{L} \cap \mathcal{D} \subset \mathcal{S}, \tag{3}$$

see e.g. [10], pp. 33 and 54.

In [9], we treated the case $s = 2$ in detail and found the *exact asymptotics* for $\mathbb{P}\{D > x\}$. We also described the *most probable way for the occurrence of the large deviations*. That means that, for

the stationary waiting time to be large, two large service times have to be large if $\rho < 1$ and B_r is a subexponential distribution (see [9, Theorem 1]) and one service time has to be large if $1 < \rho < 2$ and if B is *long-tailed* and B_r is *intermediate regularly varying* (see [9, Theorem 2]). We also obtained a number of simple bounds. First, Theorem 1 in [9] yields the following

THEOREM 1.1 *Let $s = 2$, $\rho < 1$, and let the residual time distribution B_r be subexponential. Then the tail of the stationary waiting time satisfies the asymptotic relation, as $x \to \infty$,*

$$\mathbb{P}\{D > x\} \quad \sim \quad \frac{\rho^2}{2-\rho}\Big[(\overline{B}_r(x))^2 + \int_0^\infty \overline{B}_r(x+ya)\overline{B}(x+y(a-b))dy\Big].$$

As a corollary, one can obtain the following bounds for the stationary waiting time, as $x \to \infty$:

$$\Big(\frac{\rho^2(2+\rho)}{2(2-\rho)} + o(1)\Big)\overline{B}_r^2(x) \le \mathbb{P}\{D > x\} \le \Big(\frac{\rho^2}{2(1-\rho)} + o(1)\Big)\overline{B}_r^2(x).$$

Another corollary is: if, in addition, the distribution B is regularly varying with index $\gamma > 1$, then, as $x \to \infty$:

$$\mathbb{P}\{D > x\} \quad \sim \quad c(\overline{B}_r(x))^2,$$

where

$$c \quad = \quad \frac{\rho^2}{2-\rho}\Big[1 + \frac{\rho}{\gamma-1}\int_0^\infty \frac{dz}{(1+z)^{\gamma-1}(1+z(1-\rho))^\gamma}\Big].$$

For the case $\rho > 1$, we also proved in [9]

THEOREM 1.2 *Let $s = 2$, $1 < \rho < 2$, and let both B and B_r be subexponential distributions. Then the tail of the stationary waiting time satisfies the following inequalities:*

$$\limsup_{x\to\infty} \frac{\mathbb{P}\{D > x\}}{\overline{B}_r(2x)} \quad \le \quad \frac{\rho}{2-\rho},$$

and, for any fixed $\delta > 0$,

$$\liminf_{x\to\infty} \frac{\mathbb{P}\{D > x\}}{\overline{B}_r\big(\frac{\rho+\delta}{\rho-1}x\big)} \quad \ge \quad \frac{\rho}{2-\rho}.$$

If, in particular, B is subexponential and B_r is intermediate regularly varying, then

$$\mathbb{P}\{D > x\} \quad \sim \quad \frac{\rho}{2-\rho}\overline{B}_r\Big(\frac{\rho}{\rho-1}x\Big) \qquad \text{as } x \to \infty.$$

For an arbitrary $s \ge 2$ number of servers, the best result on the existence of moments was obtained in [17, Theorem 4.1] (here $\mathcal{L}_1^\gamma$ is a specific class of distributions introduced in [17]):

THEOREM 1.3 *Let $k < \rho < k+1$ for some $k \in \{0, 1, \dots, s-1\}$. Then:*

(i) If $\mathbb{E}\sigma^\gamma < \infty$ then $\mathbb{E}D^{(s-k)(\gamma-1)} < \infty$.

(ii) If in addition σ is in the class $\mathcal{L}_1^\gamma$, then $\mathbb{E}D^{(s-k)(\gamma-1)} < \infty$ implies $\mathbb{E}S^\gamma < \infty$.

In the present paper we introduce a condition which is both necessary and sufficient for the finiteness of $\mathbb{E}D^\gamma$. We present this condition in "probabilistic terms".

THEOREM 1.4 *Let $\sigma_{r,1}$, $\sigma_{r,2}$, ... be independent random variables with common distribution B_r. Let $k < \rho < k+1$ for some $k \in \{0, 1, \dots, s-1\}$. For any $\gamma > 0$, $\mathbb{E}D^\gamma$ is finite if and only if*

$$\mathbb{E}(\min(\sigma_{r,1}, \dots, \sigma_{r,s-k}))^\gamma < \infty, \tag{4}$$

see Section 8 for the proof. Actually, this result (which is sharper than Theorem 1.3) may be deduced from the results of [17], but was not stated there. The corresponding proof in [17] involves a comparison with the so-called semi-cyclic service discipline. To the best of our knowledge, the latter approach does not allow one to obtain upper bounds for the tail distribution of D.

The main aim of the present paper is to introduce a novel approach for constructing upper bounds for the stationary waiting time in multi-server queues (see Section 7 below). This allows us to derive estimates for the tail probabilities of the distribution of the stationary waiting time if the common distribution of service times is of supexponential type, and, further, to establish the *principle of big jumps* in a particular case of intermediate varying distributions. Also, based on the new approach, we will obtain a direct proof of Theorem 1.4 (see Section 8).

The most explicit bounds are obtained for the case $\rho < 1$.

THEOREM 1.5 *Let $\rho = b/a < 1$ and let the residual time distribution B_r be subexponential. Then the tail distribution of the stationary waiting time admits the following bounds:*

$$\frac{\rho^s}{s!} \le \liminf_{x\to\infty} \frac{\mathbb{P}\{D > x\}}{\overline{B}_r^s(x)} \quad \le \quad \limsup_{x\to\infty} \frac{\mathbb{P}\{D > x\}}{\overline{B}_r^s(x)} \le \Big(\frac{\rho}{1-\rho}\Big)^s.$$

We present here the lower and upper bounds only. As it was described in [9], the only case where the tail asymptotics are available with an explicit constant multiplier is the case of regularly varying service time distribution. The corresponding calculations are rather involved and deal with the law of large numbers and a summation over a specific s-dimensional domain with planar boundaries. These calculations have been carried out in [9] in the case of $s = 2$ servers.

The proof of Theorem 1.5 (see Section 3) is based on a simple argument which cannot be applied if $\rho > 1$. For an arbitrary ρ, we have the following result.

THEOREM 1.6 *Let $k \in \{0, \dots, s-1\}$ and $\delta > 0$. If $\rho > k$, then*

$$\mathbb{P}\{D > x\} \quad \ge \quad \frac{\rho^{s-k} + o(1)}{(s-k)!} \overline{B_r}^{s-k}\Big(\frac{\rho+\delta}{\rho-k}x\Big) \quad \text{as } x \to \infty.$$

If $\rho < k+1$ and if the residual service time distribution B_r is subexponential, then

$$\mathbb{P}\{D > x\} \quad \le \quad \binom{s}{k}\Big(\frac{(k+1)\rho}{k+1-\rho} + o(1)\Big)^{s-k} \overline{B}_r^{s-k}(x(1-\delta)) \quad \text{as } x \to \infty.$$

The lower bound follows from Theorem 5.1 in Section 5. The proof of the upper bound may be found in Section 7. It is based on results from Section 6, where we present a novel construction of a consistent majorant for D_n.

Note that the lower and the upper bounds in Theorem 1.6 are not necessarily of the same order. In particular, if distribution B is of Weibull type then the ratio of the upper and the lower bounds tends to infinity, as x increases. In this case we do not have any ideas about how correct/exact/sharp bounds would look like. But if, in particular, the residual service time distribution belongs to the class $\mathcal{D} \cap \mathcal{L}$, then these bounds differ by a multiplicative constant only.

Note that in Theorem 1.6 we require conditions on the residual distribution B_r and not on the distribion B itself. This is in line with the key results on subexponentiality like (2).

COROLLARY 1.1 *Let the residual service time distribution B_r be long-tailed and dominated varying. Let $k < \rho < k+1$ for some $k \in \{0, 1, \dots, s-1\}$. Then there exist constants c_1 and c_2 such that, for all x,*

$$c_1 \overline{B}_r^{s-k}(x) \le \mathbb{P}\{D > x\} \le c_2 \overline{B}_r^{s-k}(x). \tag{5}$$

The result follows directly from Theorem 1.6, the last inclusion in (3) and the definition of the dominated variation.

In the particular case of distributions of intermediate variation, we will use [3, Theorem 7] to complement Corollary 1.1 by establishing the "principle of $s-k$ big jumps": the main cause of the value of D to be big is to have $s-k$ big service times, see Section 7 for the precise statement.

2. Comparison of systems with different inter-arrival times. Here we present results which, in particular, allow us to obtain lower and upper bounds for the stationary delay in a general $GI/GI/s$ system in terms of a simpler $D/GI/s$ system with deterministic interarrival times. We use the following partial ordering: for two vectors $\mathbf{x} = (x_1, \dots, x_s)$ and $\mathbf{y} = (y_1, \dots, y_s)$, we write $\mathbf{x} \le \mathbf{y}$ if $x_j \le y_j$ for all $j = 1, \dots, s$.

Consider two $GI/GI/s$ systems, say $\widetilde{\mathbf{V}}$ and $\widehat{\mathbf{V}}$, with service times σ_n and with interarrival times $\widetilde{\tau}_n$ and $\widehat{\tau}_n$ respectively. Let $\widetilde{D}_n$ and $\widehat{D}_n$ be the corresponding waiting times in these systems. Let $\xi_n = \widehat{\tau}_{n+1} - \widetilde{\tau}_{n+1}$. We obtain an upper bound for delay $\widetilde{D}_n$ in terms of delay $\widehat{D}_n$ and the sequence ξ_n.

LEMMA 2.1 *For all $n \ge 1$, $\widetilde{D}_n \le \widehat{D}_n + M_{n-1}$, where $M_0 = 0$ and $M_n = (M_{n-1} + \xi_n)^+$.*

PROOF. Put $\mathbf{e}_1 = (1, 0, \dots, 0)$ and $\mathbf{1} = (1, \dots, 1)$. It suffices to prove the inequality

$$\widetilde{\mathbf{W}}_n \le \widehat{\mathbf{W}}_n + \mathbf{1}M_{n-1} \quad \text{a.s.} \tag{6}$$

We proceed by induction. For $n = 1$ we have $\mathbf{0} \le \mathbf{0} + \mathbf{1}M_0$. Assume inequality (6) to hold for some n and prove it for $n + 1$. We have

$$\begin{aligned} \widetilde{\mathbf{W}}_{n+1} &= R(\widetilde{\mathbf{W}}_n + \mathbf{e}_1\sigma_n - \mathbf{1}\widetilde{\tau}_{n+1})^+ \\ &\le R(\widehat{\mathbf{W}}_n + \mathbf{1}M_{n-1} + \mathbf{e}_1\sigma_n - \mathbf{1}\widetilde{\tau}_{n+1})^+ \\ &= R(\widehat{\mathbf{W}}_n + \mathbf{e}_1\sigma_n - \mathbf{1}\widehat{\tau}_{n+1} + \mathbf{1}(M_{n-1} + \xi_n))^+. \end{aligned}$$

Since $(u + v)^+ \le u^+ + v^+$,

$$\widetilde{\mathbf{W}}_{n+1} \le R(\widehat{\mathbf{W}}_n + \mathbf{e}_1\sigma_n - \mathbf{1}\widehat{\tau}_{n+1})^+ + \mathbf{1}(M_{n-1} + \xi_n)^+ \equiv \widehat{\mathbf{W}}_{n+1} + \mathbf{1}M_n,$$

and the proof of (6) is complete. □

The following corollary will be used to obtain lower bounds. It is similar to Lemma 2 in [9].

COROLLARY 2.1 *Let $\mathbf{W}'_n$ be a stable s-server queue system with the same service times σ_n as in $\mathbf{W}_n$ and with the constant interarrival times a'. If $a' > a = \mathbb{E}\tau$, then, for any $\varepsilon > 0$, there exists x_0 such that*

$$\mathbb{P}\{D > x\} \ge (1 - \varepsilon)\mathbb{P}\{D' > x + x_0\} \quad \textit{for all } x.$$

One can take x_0 such that

$$\mathbb{P}\Big\{\sup_{n \ge 0} \sum_{i=1}^{n} (\tau_i - a') \le x_0\Big\} \ge 1 - \varepsilon.$$

PROOF. Take $\widetilde{\tau}_n = a'$ and $\widehat{\tau}_n = \tau_n$ in Lemma 2.1, then $\xi_n = \tau_n - a'$. A weak limit, M, of the sequence M_n exists (since $\mathbb{E}\xi_1 = a - a' < 0$) and has the same distribution as

$$M =_{\text{st}} \max\{0,\ \xi_1,\ \xi_1 + \xi_2,\ \dots,\ \xi_1 + \dots + \xi_n,\ \dots\}.$$

By Lemma 2.1, $D'_n \le D_n + M_{n-1}$. Hence, $D_n \ge D'_n - M_{n-1}$ Since D'_n does not depend on τ's, D'_n and M_{n-1} are independent. Therefore,

$$\mathbb{P}\{D_n > x\} \ge \mathbb{P}\{M_{n-1} \le x_0\}\mathbb{P}\{D'_n > x + x_0\}.$$

Letting n go to infinity, we obtain the desired bound. □

3. The case $\rho < 1$, proof of Theorem 1.5. The lower bound in Theorem 1.5 follows from Lemma 3.1 below which also generalises Theorem 5.1 (see Section 5) in the case $k = 0$.

LEMMA 3.1 *Let $\rho > 0$. Then, for any function $h(x) \to \infty$ as $x \to \infty$,*

$$\mathbb{P}\{D > x\} \ge \frac{\rho^s + o(1)}{s!}\overline{B_r}^s(x + h(x)).$$

In particular, if the residual time distribution B_r is long-tailed (that is, $\overline{B}_r(x+1) \sim \overline{B}_r(x)$ as $x \to \infty$), then

$$\mathbb{P}\{D > x\} \ge \frac{\rho^s + o(1)}{s!}\overline{B_r}^s(x) \quad \textit{as } x \to \infty.$$

We start with an auxiliary result.

LEMMA 3.2 *Let $\{q_i\}_{i\geq 1}$ be a non-increasing sequence of positive numbers. Then, for any $s \geq 1$,*

$$\sum_{1\leq i_1<\ldots<i_s} q_{i_1}\cdot\ldots\cdot q_{i_s} \quad \geq \quad \frac{1}{s!}(q_s+q_{s+1}+\ldots)^s.$$

PROOF. If $1 \leq i_1 < \ldots < i_s$ then $s \leq i_1+(s-1) \leq i_2+(s-2) \leq \ldots \leq i_{s-1}+1 \leq i_s$ and

$$q_{i_1}\cdot q_{i_2}\cdot\ldots\cdot q_{i_s} \geq q_{i_1+s-1}\cdot q_{i_2+s-2}\cdot\ldots\cdot q_{i_s},$$

because $\{q_i\}$ is a non-increasing sequence. Thus,

$$\begin{aligned}\sum_{1\leq i_1<\ldots<i_s} q_{i_1}\cdot\ldots\cdot q_{i_s} &\geq \sum_{s\leq i_1\leq\ldots\leq i_s} q_{i_1}\cdot\ldots\cdot q_{i_s}\\ &\geq \frac{1}{s!}\sum_{i_1,\ldots,i_s\geq s} q_{i_1}\cdot\ldots\cdot q_{i_s},\end{aligned}$$

which yields the conclusion of the lemma. □

PROOF OF LEMMA 3.1. Our estimation is based on calculations involving s big jumps. This technique was already used in [9] in the case $s = 2$, where a lower bound (which is better than the one presented in Lemma 3.1) was obtained under the extra condition that B_r is long-tailed. The bound in [9] is exact in the sense that it provides the right asymptotics under further assumptions.

Following Corollary 2.1, define the auxiliary s-server system $\mathbf{W}'_n$ having the same service times σ_n and constant interarrival times a', $a' > a$. For $\mathbf{i} = (i_1,\ldots,i_s)$, $1 \leq i_1 < \ldots < i_s < n$, define events $A_n(\mathbf{i})$ and $C_n(\mathbf{i})$ as

$$A_n(\mathbf{i}) \quad = \quad \{\sigma_{i_1} > x+(n-i_1)a', \ldots, \sigma_{i_s} > x+(n-i_s)a'\}$$

and

$$C_n(\mathbf{i}) \quad = \quad \bigcap_{i\leq n, i\neq i_1,\ldots,i_s} \{\sigma_i \leq x+(n-i)a'\}.$$

Since the mean $\mathbb{E}\sigma$ exists, we have that, uniformly in n and $\mathbf{i}$,

$$\mathbb{P}\{C_n(\mathbf{i})\} = 1-\mathbb{P}\{\overline{C_n(\mathbf{i})}\} \quad \geq \quad 1-\sum_{i=0}^{\infty}\mathbb{P}\{\sigma_1 > x+ia'\} \to 1 \quad \text{as } x\to\infty.$$

For each vector $\mathbf{i}$, events $A_n(\mathbf{i})$ and $C_n(\mathbf{i})$ are independent. Further, events $A_n(\mathbf{i})\cap C_n(\mathbf{i})$ are disjoint for distinct vectors $\mathbf{i}$. These observations together yield

$$\mathbb{P}\Big\{\bigcup_{\mathbf{i}} A_n(\mathbf{i})\cap C_n(\mathbf{i})\Big\} \quad = \quad \sum_{\mathbf{i}}\mathbb{P}\{A_n(\mathbf{i})\}\mathbb{P}\{C_n(\mathbf{i})\} \geq (1-o(1))\sum_{\mathbf{i}}\mathbb{P}\{A_n(\mathbf{i})\} \tag{7}$$

as $x\to\infty$, uniformly in n. The event $A_n(\mathbf{i})$ implies that $D'_n > x$. Therefore,

$$\mathbb{P}\{D'_n > x\} \quad \geq \quad (1-o(1))\sum_{\mathbf{i}}\mathbb{P}\{A_n(\mathbf{i})\}$$

as $x\to\infty$, uniformly in n. We now prove that

$$\lim_{n\to\infty}\sum_{\mathbf{i}}\mathbb{P}\{A_n(\mathbf{i})\} \quad \geq \quad \frac{(b/a')^s}{s!}\overline{B_r}^s(x+sa'). \tag{8}$$

Indeed, by the independence of the σ's,

$$\sum_{1\leq i_1<\ldots<i_s<n}\mathbb{P}\{A_n(\mathbf{i})\} \quad = \quad \sum_{1\leq i_1<\ldots,i_s<n}\overline{B}(x+(n-i_1)a')\cdot\ldots\cdot\overline{B}(x+(n-i_s)a'),$$

and the left side of (8) equals

$$\sum_{1\leq i_1<\ldots<i_s}\overline{B}(x+i_1a')\cdot\ldots\cdot\overline{B}(x+i_sa').$$

By Lemma 3.2 with $q_i = \overline{B}(x + ia')$, the latter sum is not smaller than

$$\frac{1}{s!}\Big(\sum_{j=s}^{\infty} \overline{B}(x + ja')\Big)^s.$$

Since the tail probability is a non-increasing function,

$$\sum_{j=s}^{\infty} \overline{B}(x + ja') \geq \frac{1}{a'} \int_{sa'}^{\infty} \overline{B}(x + z)dz = \rho \overline{B}_r(x + sa').$$

Combining altogether, we conclude (8). Then by Corollary 2.1, for every $\varepsilon > 0$ there exists x_0 such that

$$\begin{aligned} \mathbb{P}\{D > x\} &\geq (1 - \varepsilon)\mathbb{P}\{D' > x + x_0\} \\ &\geq (1 - \varepsilon - o(1))\frac{(b/a')^s}{s!}\overline{B_r}^s(x + sa' + x_0). \end{aligned}$$

By the arbitrary choice of $a' > a$ and $\varepsilon > 0$, the proof of Lemma 3.1 is complete. □

Proof of the upper bound in Theorem 1.5. We start with the case of deterministic τ, i.e., $\tau_n \equiv a$. We follow the lines from [9] where, for $\rho < 1$ (that is for $b < a$), the following simple majorant was introduced.

Let σ_{ni}, $n \geq 1$, $i \leq s$, be independent random variables with common distribution B. Consider s auxiliary $D/GI/1$ queueing systems which work in parallel: at every time instant $T_n = na$, $n = 1, 2, \ldots$, a batch of s customers arrives, one customer per each queue. Service times in queue i are equal to σ_{ni}. Denote by U_{ni}, $i = 1, \ldots, s$, the waiting times in the ith queue, $U_{n+1,i} = (U_{ni} + \sigma_{ni} - a)^+$, and let $U_{1i} = 0$. Since the arrival process is deterministic and service times are independent, vector $(U_{n1}, \ldots, U_{ns})$ has independent identically distributed coordinates, and, as $n \to \infty$, its weak limit $(U_1, \ldots, U_s)$ exists (since $\mathbb{E}\sigma < a$) and contains independent identically distributed coordinates too. Here U_i is the stationary waiting time in the ith auxiliary queue.

Now we introduce a coupling of s single-server systems and of the s-server system $D/GI/s$. Namely, we determine the service times σ_n in the original $D/GI/s$ system by induction. Start with $\sigma_1 = \sigma_{1,1}$. Assume that $\sigma_1, \ldots, \sigma_{n-1}$ have been already defined. Then the delay vectors $\mathbf{V}_1, \ldots, \mathbf{V}_n$ are defined too, and we know the number $i_n = \min\{i : V_{ni} = D_n\}$. Then let $\sigma_n = \sigma_{n,i_n}$.

By monotonicity, $D_n \leq \min\{U_{ni},\ i \leq s\}$ with probability 1. Hence,

$$D \leq \min\{U_i,\ i \leq s\}. \tag{9}$$

Due to independence,

$$\mathbb{P}\{D > x\} \leq \mathbb{P}^s\{U_1 > x\},$$

and we can apply known results for the single server queue: from (2),

$$\mathbb{P}\{U_1 > x\} \sim \frac{\rho}{1 - \rho}\overline{B}_r(x),$$

which gives us the upper bound in Theorem 1.5 if interarrival times are deterministic. Now the proof in the general case follows from [9, Lemma 1]. □

4. Auxiliary results. In this Section we collect a number of auxiliary facts related to monotonicity and to the strong law of large numbers for unstable multi-server systems. The results seem not to be new, so we provide only short sketches of proofs for self-containedness.

Let $\mathbf{W}_n$ be a sequence satisfying the Kiefer-Wolfowitz recursion (1), with initial value $\mathbf{W}_1 \geq 0$.

Lemma 4.1 *(1) For any n, $\mathbf{W}_n$ is a non-decreasing function of the initial value and of service times and a non-increasing function of interarrival times. This means that if $\widetilde{\mathbf{W}}_n$ is a sequence satisfying another Kiefer-Wolfowitz recursion with initial value $\widetilde{\mathbf{W}}_1$ and with interarrival times $\{\widetilde{\tau}_n\}$ and service times $\{\widetilde{\sigma}_n\}$ and if $\mathbf{W}_1 \leq \widetilde{\mathbf{W}}_1$ (coordinate-wise), $\sigma_j \leq \widetilde{\sigma}_j$, and $\tau_j \geq \widetilde{\tau}_j$, for $j = 1, \ldots, n-1$, then $\mathbf{W}_n \leq \widetilde{\mathbf{W}}_n$.*
(2) For any $n \geq 2$, the difference $\sum_{i=1}^s (W_{ni} - W_{n-1,i})$ is a non-increasing function of the initial value $\mathbf{W}_1$: if $\mathbf{W}_1 \leq \widetilde{\mathbf{W}}_1$, then $\sum_{i=1}^s (W_{ni} - W_{n-1,i}) \geq \sum_{i=1}^s (\widetilde{W}_{ni} - \widetilde{W}_{n-1,i})$.

The first monotonicity property holds because both operators R and $\max(0,\cdot)$ are monotone. The second property follows since function $(x+y)^+ - x$ is non-increasing in x, for any fixed y.

LEMMA 4.2 *Let $b > sa$, so the s-server system with workload vectors $\mathbf{W}_n$ is unstable. Then,*

$$\frac{W_{n1}}{n} \to \frac{b-sa}{s} \quad \textit{and} \quad \frac{W_{ns}}{n} \to \frac{b-sa}{s} \quad \textit{as } n \to \infty, \tag{10}$$

both with probability 1 *and in mean.*

PROOF. Note that, for any $n = 1, 2, \ldots$,

$$W_{n+1,s} - W_{n+1,1} \le \max(W_{1,s} - W_{1,1}, \sigma_1, \ldots, \sigma_n). \tag{11}$$

Indeed, if $W_{ns} - W_{n1} > \sigma_n$, then $W_{n+1,s} - W_{n+1,1} \le W_{ns} - W_{n1}$, and if $W_{ns} - W_{n1} \le \sigma_n$, then $W_{n+1,s} - W_{n+1,1} \le \sigma_n$, so the induction argument completes the proof of (11). Next,

$$\max(W_{1s} - W_{11}, \sigma_1, \ldots, \sigma_n)/n \to 0 \quad \text{a.s.} \tag{12}$$

because $(W_{1s} - W_{11})/n \to 0$ and, since $\mathbf{E}\sigma$ is finite, events $\{\sigma_k/k > \varepsilon\}$ occur only finitely often, for any $\varepsilon > 0$.

Further,

$$\frac{1}{n}\sum_{i=1}^{s} W_{ni} \ge \frac{1}{n}\sum_{j=1}^{n-1}(\sigma_j - s\tau_{j+1}) \to b - sa > 0 \quad \text{a.s.},$$

so $\liminf_{n\to\infty} W_{ns}/n \ge (b-sa)/s$, and, from (11)-(12), there exists an a.s. finite random variable ν such that $W_{n1} > 0$, for all $n \ge \nu$. So, for $n \ge \nu$,

$$\frac{1}{n}\sum_{i=1}^{s} W_{ni} = \frac{1}{n}\sum_{i=1}^{s} W_{\nu i} + \frac{1}{n}\sum_{j=\nu}^{n}(\sigma_j - s\tau_{j+1}) \to b - sa \quad \text{a.s.}, \tag{13}$$

and (11)-(13) lead to convergence a.s. in (10). Finally, since $0 \le W_{ns}/n \le W_{1s}/n + \sum_{j=1}^{n-1}\sigma_j/n$ and since random variables $\sum_{j=1}^{n-1}\sigma_j/n$ are uniformly integrable, convergence in mean also follows. □

LEMMA 4.3 *Assume $b > (s-1)a$. For any $\varepsilon > 0$, there exist $A < \infty$ and an integer $d \ge 1$ such that, for any initial value $\mathbf{W}_1$ with $W_{1s} \ge A$,*

$$\mathbb{E}\{W_{1+d,1} + \ldots + W_{1+d,s} - W_{11} - \ldots - W_{1s}\} \le d(b - sa + \varepsilon).$$

PROOF. By property (2) of Lemma 4.1, it is enough to prove the result for initial value $W_{11} = \ldots = W_{s-1,1} = 0$, $W_{s1} = A$ only.

Choose C such that $\mathbb{E}\min(\tau, C) \ge a - \varepsilon/2$. By property (1) of Lemma 4.1, we may prove the lemma with interarrival times $\min(\tau_j, C)$ in place of τ_j.

Consider an auxiliary unstable $GI/GI/(s-1)$ queue $\widehat{\mathbf{W}}_n$ with initial zero value and, by applying the previous lemma, find d such that $\mathbb{E}\sum_{i=1}^{s-1}\widehat{W}_{1+d,i} \le d(b - (s-1)a + \varepsilon/2)$. Then return to the s-server queue and take $A = (d+1)C$. We will prove that

$$\sum_{i=1}^{s} W_{1+d,i} = \sum_{i=1}^{s-1}\widehat{W}_{1+d,i} + A - \sum_{j=1}^{d}\min(\tau_j, C) \quad \text{a.s.}, \tag{14}$$

then the result will follow.

Consider vectors $\mathbf{V}_n$ and numbers i_n as in the Introduction, with initial values $V_{1,1} = \ldots = V_{1,s-1} = 0$ and $V_{1,s} = A$. Note that $V_{n,s} \ge A - (n-1)C > 0$, for all $n = 1, 2, \ldots, d+1$.

Let $\mu = \min(d+1, \min\{n \ge 1 \ : \ i_n = s\})$. Then $R(V_{\mu,1}, \ldots, V_{\mu,s-1}) = (\widehat{W}_{\mu,1} \ldots, \widehat{W}_{\mu,s-1})$ and

$$\sum_{i=1}^{s} W_{\mu,i} = \sum_{i=1}^{s} V_{\mu,i} = \sum_{i=1}^{s-1} V_{\mu,i} + V_{\mu,s} = \sum_{i=1}^{s-1}\widehat{W}_{\mu,i} + A - \sum_{j=1}^{\mu-1}\min(\tau_j, C). \tag{15}$$

This ends the proof of (14) if $\mu = d+1$.

In the case $\mu < d+1$, we may conclude that

$$0 < A - (\mu-1)C \le V_{\mu,s} = W_{\mu,1}$$

and, therefore, $W_{n,i} > 0$ and $\widehat{W}_{n,i} > 0$, for all $\mu \le n \le d+1$ and $i = 1, \dots, s$. Then, from (15),

$$\begin{aligned}\sum_{i=1}^{s} W_{d+1,i} &= \sum_{i=1}^{s} W_{\mu,i} + \sum_{j=\mu}^{d} (\sigma_j - s\min(\tau_j, C)) \\ &= \sum_{i=1}^{s-1} \widehat{W}_{\mu,i} + \sum_{j=\mu}^{d} (\sigma_j - (s-1)\min(\tau_j, C)) + A - \sum_{j=1}^{d} \min(\tau_j, C)\end{aligned}$$

which coincides again with the right side of (14).

□

5. Lower Bound. The following result holds without any restrictions on the service time distribution B (a similar result was formulated and proved in [17, Theorem 3.1]).

THEOREM 5.1 *Let $k \in \{0, 1, \dots, s-1\}$ be such that $\rho > k$. Then, for any fixed $\delta > 0$,*

$$\mathbb{P}\{D > x\} \ge \frac{\rho^{s-k} + o(1)}{(s-k)!} \overline{B_r}^{s-k}\Big(\frac{\rho+\delta}{\rho-k}x\Big) \quad \text{as } x \to \infty.$$

PROOF. We exploit the technique of $s-k$ big jumps. Following Corollary 2.1, we consider only deterministic interarrival times, $\tau \equiv a$.

The case $k = 0$ was considered in Lemma 3.1. So now let $k \ge 1$. Let $\widetilde{\mathbf{W}}_n = (\widetilde{W}_{n1}, \dots, \widetilde{W}_{nk})$ be the residual workload vector in the $GI/GI/k$ system with the same interarrival and service times as in the original system, and with k servers. Since $\rho > k$, the k-server system is unstable. Hence, by Lemma 4.2, both the minimal coordinate $\widetilde{W}_{n1}$ and the maximal coordinate $\widetilde{W}_{nk}$ drift to infinity as $n \to \infty$ with probability 1, with the same rate $(b-ka)/k$. Then

$$\mathbb{P}\Big\{\widetilde{W}_{N1} > N\Big(\frac{b-ka}{k} - \delta\Big),\ \widetilde{W}_{ik} \le N\Big(\frac{b-ka}{k} + \delta\Big) \text{ for all } i \le N\Big\} \quad \to \quad 1 \ \text{ as } N \to \infty.$$

If we assume that there are initially big workloads at $s-k$ servers while the k other queues are empty, then, with high probability, the k smallest workloads evolve like the k-server system with workloads $\widetilde{\mathbf{W}}_n$, for a long while. This observation implies that

$$\begin{aligned}\mathbb{P}\Big\{W_{N1} > N\Big(\frac{b-ka}{k} - \delta\Big),\ W_{ik} \le N\Big(\frac{b-ka}{k} + \delta\Big), W_{i,k+1} > N\Big(\frac{b-ka}{k} + \delta\Big) \text{ for all } i \le N \Big| \\ W_{1k} = 0, W_{1,k+1} > N\Big(\frac{b-ka}{k} + \delta\Big) + Na\Big\} \to 1 \ \text{ as } N \to \infty.\end{aligned}$$

Take c such that

$$\frac{b-ka}{k} + \delta \quad \le \quad (1+c\delta)\Big(\frac{b-ka}{k} - \delta\Big) \tag{16}$$

for all sufficiently small $\delta > 0$, and let

$$x = N\Big(\frac{b-ka}{k} - \delta\Big). \tag{17}$$

Then

$$\mathbb{P}\{D_N > x \mid W_{1k} = 0, W_{1,k+1} > x(1+c\delta) + Na\} \quad \to \quad 1 \ \text{ as } x \to \infty.$$

By the monotonicity of the s-server queueing system in its initial state (see Lemma 4.1), we obtain

$$\mathbb{P}\{D_N > x \mid W_{1,k+1} > x(1+c\delta) + Na\} \quad \to \quad 1 \ \text{ as } x \to \infty. \tag{18}$$

For $\mathbf{i} = (i_1, \dots, i_{s-k})$, $1 \le i_1 < \dots < i_{s-k} \le n$, define the events $A_n(\mathbf{i})$ as

$$A_n(\mathbf{i}) \quad = \quad \{\sigma_{i_1} > y + (n-i_1)a, \dots, \sigma_{i_{s-k}} > y + (n-i_{s-k})a\}.$$

Again like in (7) we have

$$\mathbb{P}\Big\{\bigcup_{\mathbf{i}:i_{s-k}<n-N} A_n(\mathbf{i})\Big\} \;\geq\; (1-o(1))\sum_{\mathbf{i}:i_{s-k}<n-N}\mathbb{P}\{A_n(\mathbf{i})\} \tag{19}$$

as $y\to\infty$, uniformly in n and N. We prove now that

$$\lim_{n\to\infty}\sum_{\mathbf{i}:i_{s-k}<n-N}\mathbb{P}\{A_n(\mathbf{i})\} \;\geq\; \frac{\rho^{s-k}}{(s-k)!}\overline{B_r}^{s-k}(y+(N+s-k)a). \tag{20}$$

Indeed, by the independence of the σ's,

$$\begin{aligned}\sum_{\mathbf{i}:i_{s-k}<n-N}\mathbb{P}\{A_n(\mathbf{i})\} &= \sum_{\mathbf{i}:i_{s-k}<n-N}\overline{B}(y+(n-i_1)a)\cdot\ldots\cdot\overline{B}(y+(n-i_{s-k})a)\\ &= \sum_{N<i_1<\ldots<i_{s-k}\leq n-1}\overline{B}(y+i_1a)\cdot\ldots\cdot\overline{B}(y+i_{s-k}a).\end{aligned}$$

Hence, the left side of (20) equals

$$\sum_{N<i_1<\ldots<i_{s-k}}\overline{B}(y+i_1a)\cdot\ldots\cdot\overline{B}(y+i_{s-k}a) \;\geq\; \sum_{1\leq i_1<\ldots<i_{s-k}}\overline{B}(y+Na+i_1a)\cdot\ldots\cdot\overline{B}(y+Na+i_{s-k}a).$$

By Lemma 3.2 with $q_i=\overline{B}(y+Na+ia)$, the sum on the right is not less than

$$\frac{1}{(s-k)!}\Big(\sum_{j=s-k}^{\infty}\overline{B}(y+Na+ja)\Big)^{s-k}.$$

Since the tail probability is non-increasing,

$$\sum_{j=s-k}^{\infty}\overline{B}(y+Na+ja)\geq\frac{1}{a}\int_{(s-k)a}^{\infty}\overline{B}(y+Na+z)dz.$$

Combining these expressions, we obtain the desired estimate (20). Substituting (20) into (19) we get, as $y\to\infty$,

$$\lim_{n\to\infty}\mathbb{P}\Big\{\bigcup_{\mathbf{i}:i_{s-k}<n-N}A_n(\mathbf{i})\Big\} \;\geq\; \frac{\rho^{s-k}-o(1)}{(s-k)!}\overline{B_r}^{s-k}(y+(N+s-k)a). \tag{21}$$

Let $y=x(1+c\delta)$ in the definition of $A_n(\mathbf{i})$. Since an increment per unit of time of every coordinate of the workload vector $\mathbf{W}_i$ is not less than $-a$, on the event $A_n(\mathbf{i})$ we have $W_{j,k+1}>x(1+c\delta)+(n-j)a$, for all $j\in[i_{s-k}+1,n]$. Consider $\mathbf{i}$ such that $i_{s-k}<n-N$. Then $A_n(\mathbf{i})$ implies $W_{n-N,k+1}>x(1+c\delta)+Na$. Together with (18) it yields

$$\mathbb{P}\Big\{D_n>x\Big|\bigcup_{\mathbf{i}:i_{s-k}<n-N}A_n(\mathbf{i})\Big\} \;\to\; 1 \tag{22}$$

as $x\to\infty$, uniformly in n and N. Now it follows from (21) and (22) that

$$\begin{aligned}\mathbb{P}\{D>x\} &\geq \liminf_{n\to\infty}\mathbb{P}\Big\{W_{n1}>x\Big|\bigcup_{\mathbf{i}:i_{s-k}<n-N}A_n(\mathbf{i})\Big\}\mathbb{P}\Big\{\bigcup_{\mathbf{i}:i_{s-k}<n-N}A_n(\mathbf{i})\Big\}\\ &\geq \frac{\rho^{s-k}-o(1)}{(s-k)!}\overline{B_r}^{s-k}(x(1+c\delta)+(N+s-k)a).\end{aligned}$$

Thus, by (17),

$$\mathbb{P}\{D>x\} \;\geq\; \frac{\rho^{s-k}-o(1)}{(s-k)!}\overline{B_r}^{s-k}\Big(x\Big(1+c\delta+\frac{ka}{b-ka-k\delta}\Big)+(s-k)a\Big).$$

Hence, for every $\delta>0$,

$$\begin{aligned}\mathbb{P}\{D>x\} &\geq \frac{\rho^{s-k}+o(1)}{(s-k)!}\overline{B_r}^{s-k}\Big(x\Big(1+\frac{ka+\delta}{b-ka}\Big)\Big)\\ &= \frac{\rho^{s-k}+o(1)}{(s-k)!}\overline{B_r}^{s-k}\Big(x\frac{b+\delta}{b-ka}\Big) \;\text{ as } x\to\infty.\end{aligned}$$

The proof of Theorem 5.1 is complete. □

6. New Majorant. In the proof of the upper bound in Theorem 1.5 (see Section 3), we introduced s parallel single server queues that provide a suitable majorant in the case $\rho < 1$. If $\rho \geq 1$ then the single server system with service time distribution B is unstable and the above scheme does not work. For an arbitrary ρ, we need a more complex procedure to obtain a majorant. Hereinafter let $k = [b/a]$ be the integer part of b/a. We continue to assume constant interarrival times $\tau \equiv a$.

Again let σ_{ni}, $n \geq 1$, $i \leq s$, be independent random variables with common distribution B. Define service times σ_n in the original $D/GI/s$ system as in Section 3. Consider again s auxiliary single server queues $D/GI/1$, but now with different deterministic arrival times $T_n = n(k+1)(a-h)$ where

$$\frac{k}{k+1}\Big(a - \frac{b}{k+1}\Big) < h < a - \frac{b}{k+1}, \tag{23}$$

and with service times equal to σ_{ni} in queue $i = 1, 2, \ldots, s$. Then $T_1 = (k+1)(a-h) > b$, so that the queues are stable. Let U_i be a stationary waiting time in the ith auxiliary queue. Since, for each n,

$$\text{the sequences } \{U_{n1}, n \geq 1\}, \ldots, \{U_{ns}, n \geq 1\} \text{ are mutually independent,} \tag{24}$$

the limiting vector $(U_1, \ldots, U_s)$ consists of independent identically distributed coordinates too.

In contrast to the case $\rho < 1$, it may not be true in general that, say, V_{n1} is smaller than U_{n1}. Nevertheless, for $\rho < k+1$, we can prove that, for any set I of $k+1$ indices, $\sum_{i\in I} V_{ni} \leq \sum_{i\in I} U_{ni} + \eta_I$ where η_I has a light-tailed distribution, this is Lemma 6.2 below. But first we state the main result of the Section which is an analogue of (9) for the general ρ.

LEMMA 6.1 *There exists a number $\beta > 0$ and a random variable η such that $\mathbb{E}e^{\beta\eta} < \infty$ and, for all n, with probability 1,*

$$D_n \quad \leq \quad U_{n,(k+1)} + \eta,$$

where $U_{n,(k+1)}$ is the $(k+1)$th order statistic of vector $(U_{n1}, \ldots, U_{ns})$.

Now we formulate and prove the following result. Based on it, we give the proof of Lemma 6.1 at the end of the section.

LEMMA 6.2 *There exists $\beta > 0$ such that, for any set of $k+1$ indices $I = \{i(1), \ldots, i(k+1)\}$, there is a random variable η_I such that $\mathbb{E}e^{\beta\eta_I} < \infty$ and, for any n, with probability 1,*

$$\sum_{i\in I} V_{ni} \quad \leq \quad \sum_{i\in I} U_{ni} + \eta_I.$$

PROOF. Fix some $i' \in I$. Consider an auxiliary $GI/GI/k+1$ system $\mathbf{V}'_n = (V'_{ni}, i \in I)$ with the same interarrival times equal to τ_n, but whose service times σ'_n are chosen in a special manner. At any time n, if $i_n \in I$ (i_n is defined in the proof of Theorem 1.5, see Section 3) then put $\sigma'_n = \sigma_{n,i_n}$ and $i'_n = i_n$. If $i_n \notin I$ then put $\sigma'_n = \sigma_{ni'}$ and $i'_n = i'$. Applying property (1) of Lemma 4.1, we get that $R(V_{ni}, i \in I) \leq R\mathbf{V}'_n$ coordinatewise, for any n. Therefore,

$$\sum_{i\in I} V_{ni} \quad \leq \quad \sum_{i\in I} V'_{ni}.$$

Hence, it suffices to prove that

$$\sum_{i\in I} V'_{ni} \quad \leq \quad \sum_{i\in I} U_{ni} + \eta_I. \tag{25}$$

For every $i \in I$ and for any n,

$$U_{n+1,i} - U_{ni} \quad \geq \quad \sigma_{ni} - (k+1)(a-h),$$

and hence

$$\begin{aligned}\sum_{i\in I} U_{n+1,i} - \sum_{i\in I} U_{ni} \quad &\geq \quad \sigma'_n - (k+1)(a-h) + \sum_{i\in I, i\neq i'_n} (\sigma_{in} - (k+1)(a-h)) \\ &= \quad \sigma'_n - (k+1)a + \zeta_{I,n},\end{aligned} \tag{26}$$

where the independent identically distributed random variables

$$\zeta_{I,n} := \sum_{i\in I, i\neq i'_n} (\sigma_{in} - b) + (k+1)^2 h + k(b-(k+1)a)$$

have a positive mean, by the left inequality in (23).

Take $\varepsilon \in (0, \mathbb{E}\zeta_{I,n})$. Let d and A be defined by Lemma 4.3 applied to the system $(V'_{1+n,i}, i \in I)$. Consider d-skeleton of $\mathbf{V}'$, that is, the sequence $\mathbf{V}'_{1+nd}$. For every n, if the maximal coordinate of $(V'_{1+nd,i}, i \in I)$ is not bigger than A, then

$$\sum_{i\in I} V'_{1+nd+d,i} - \sum_{i\in I} V'_{1+nd,i} \quad \leq \quad \sigma'_{nd+1} + \ldots + \sigma'_{nd+d},$$

which together with (26) implies that

$$\begin{aligned}\sum_{i\in I} V'_{1+nd+d,i} - \sum_{i\in I} V'_{1+nd,i} &\leq \sum_{i\in I} U_{1+nd+d,i} - \sum_{i\in I} U_{1+nd,i} + d(k+1)a - \zeta_{I,1+nd} - \ldots - \zeta_{I,nd+d} \\ &\leq \sum_{i\in I} U_{1+nd+d,i} - \sum_{i\in I} U_{1+nd,i} + d((k+1)a+b),\end{aligned}$$

since $\zeta_{I,n} \geq -b$. To conclude, if the maximal coordinate of $(V'_{1+nd,i}, i \in I)$ is not bigger than A, then

$$\sum_{i\in I} V'_{1+nd,i} \quad \leq \quad (k+1)A,$$

so that

$$\begin{aligned}\sum_{i\in I} V'_{1+nd+d,i} &\leq \sum_{i\in I} U_{1+nd+d,i} + d((k+1)a+b) + (k+1)A \\ &:= \sum_{i\in I} U_{1+nd+d,i} + C, \end{aligned} \tag{27}$$

Further, if the maximal coordinate of $(V'_{1+nd,i}, i \in I)$ is bigger than A, then, by Lemma 4.3, we have

$$\sum_{i\in I} V'_{1+nd+d,i} - \sum_{i\in I} V'_{1+nd,i} \quad \leq \quad \theta_{I,1+nd},$$

where $\theta_{I,1+nd}$ are independent identically distributed random variables with mean $\mathbb{E}\theta_{I,1+nd} \leq d(b-(k+1)a+\varepsilon)$ and such that

$$\theta_{I,1+nd} \quad \leq \quad \sigma'_{nd+1} + \ldots + \sigma'_{nd+d}.$$

In this case, by (26),

$$\begin{aligned}\sum_{i\in I} V'_{1+nd+d,i} - \sum_{i\in I} V'_{1+nd,i} &\leq \sum_{i\in I} U_{1+nd+d,i} - \sum_{i\in I} U_{1+nd,i} + \theta_{I,1+nd} - (\sigma'_{nd} + \ldots + \sigma'_{nd+d}) \\ &\qquad + d(k+1)a - \zeta_{I,1+nd} - \ldots - \zeta_{I,nd+d} \\ &:= \sum_{i\in I} U_{1+nd+d,i} - \sum_{i\in I} U_{1+nd,i} + \widetilde{\theta}_{I,1+nd}. \end{aligned} \tag{28}$$

Inequalities (27) and (28) imply that always

$$\sum_{i\in I} V'_{1+nd+d,i} \quad \leq \quad \sum_{i\in I} U_{1+nd+d,i} + C + \max_n \sum_{j=0}^{n} \widetilde{\theta}_{I,1+jd}. \tag{29}$$

Here $\widetilde{\theta}_{I,1+nd}, n = 1, 2, \ldots$ are independent identically distributed random variables that are bounded from above and have a negative mean. Indeed,

$$\begin{aligned}\widetilde{\theta}_{I,1+nd} &= \theta_{I,1+nd} - (\sigma'_{nd+1} + \ldots + \sigma'_{nd+d}) + d(k+1)a - \zeta_{I,1+nd} - \ldots - \zeta_{I,nd+d} \\ &\leq d(k+1)a - \zeta_{I,1+nd} - \ldots - \zeta_{I,nd+d} \\ &\leq d((k+1)a+b)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\widetilde{\theta}_{I,1+nd} &= \mathbb{E}\theta_{I,1+nd} - bd + d(k+1)a - d\mathbb{E}\zeta_{I,1+nd}\\ &\le d(b-(k+1)a+\varepsilon) - bd + d(k+1)a - d\mathbb{E}\zeta_{I,1+nd}\\ &= d(\varepsilon - \mathbb{E}\zeta_{I,1+nd}) < 0,\end{aligned}$$

by the choice of ε. Therefore, there exists $\beta > 0$ such that $\mathbb{E}e^{\beta\widetilde{\theta}_{I,1+nd}} < 1$ and then the following estimate holds:

$$\mathbb{E}\exp\Big\{\beta \max_n \sum_{j=0}^{n} \widetilde{\theta}_{I,1+jd}\Big\} \le \sum_n (\mathbb{E}e^{\beta\widetilde{\theta}_{I,1+nd}})^n < \infty.$$

Let

$$\widetilde{\eta}_I := C + \max_n \sum_{j=0}^{n} \widetilde{\theta}_{I,1+jd},$$

then $\mathbb{E}e^{\beta\widetilde{\eta}_I} < \infty$ and, by the upper bound (29), for all n,

$$\sum_{i\in I} V'_{1+nd+d,i} \le \sum_{i\in I} U_{1+nd+d,i} + \widetilde{\eta}_I. \tag{30}$$

Since, in addition, for every $l \in [1,d]$,

$$\sum_{i\in I} V'_{1+nd+d+l,i} - \sum_{i\in I} V'_{1+nd+d+l-1,i} \le \sum_{i\in I} U_{1+nd+d+l,i} - \sum_{i\in I} U_{1+nd+d+l-1,i} + (k+1)a,$$

we conclude from (30) that, for every $l \in [1,d]$,

$$\sum_{i\in I} V'_{1+nd+d+l,i} \le \sum_{i\in I} U_{1+nd+d+l,i} + \widetilde{\eta}_I + l(k+1)a,$$

so that, for every n,

$$\sum_{i\in I} V'_{ni} \le \sum_{i\in I} U_{ni} + \widetilde{\eta}_I + d(k+1)a,$$

which completes the proof of Lemma 6.2 with $\eta_I := \widetilde{\eta}_I + d(k+1)a$. □

Proof of Lemma 6.1. For every collection I of $k+1$ coordinates

$$D_n \le \frac{1}{k+1}\sum_{i\in I} V_{ni},$$

since D_n is the minimal coordinate. Then it follows from Lemma 6.2 that

$$D_n \le \frac{1}{k+1}\sum_{i\in I} U_{ni} + \eta_I. \tag{31}$$

Take $\eta := \max_{I:|I|=k+1} \eta_I$. Then $\mathbb{E}e^{\beta\eta} < \infty$ and

$$D_n \le \frac{1}{k+1}\sum_{i\in I} U_{ni} + \eta. \tag{32}$$

Take I such that $\{U_{ni}, i \in I\}$ are the $k+1$ smallest coordinates of vector $(U_{n1},\dots,U_{ns})$. Then $U_{ni} \le U_{n,(k+1)}$ for every $i \in I$. Together with (32) it yields the inequality of the lemma. □

7. Upper Bound and the Principle of $s-k$ Big Jumps. Now we turn to the upper bound. Lemma 6.1 allows to prove the following general result.

Theorem 7.1 *Let $\rho < k+1$ for some $k \in \{0,1,\dots,s-1\}$. Then, for any fixed h satisfying* (23)*, there exists $\beta > 0$ such that*

$$\mathbb{P}\{D > x+y\} \le \binom{s}{k}(\overline{F}(x))^{s-k} + const\cdot e^{-\beta y} \quad \textit{for all } x,y>0,$$

where F is the distribution of random variable

$$M := \sup\Big(0,\ \sum_{j=1}^{n}(\sigma_j - (k+1)(a-h)),\ n \ge 1\Big).$$

PROOF. First, by inequality $b \equiv \mathbb{E}\sigma < (k+1)(a-h)$ and by the strong law of large numbers, the maximum M is finite with probability 1. By Lemma 6.1,

$$\begin{aligned}\mathbb{P}\{D_n > x+y\} &\leq \mathbb{P}\{U_{n,(k+1)} + \eta > x+y\} \\ &\leq \mathbb{P}\{U_{n,(k+1)} > x\} + \mathbb{P}\{\eta > y\}.\end{aligned}$$

Taking into account the independence of the U's in (24), we obtain

$$\mathbb{P}\{D_n > x+y\} \leq \left(\begin{array}{c} s \\ k \end{array} \right) \mathbb{P}^{s-k}\{U_{n1} > x\} + \mathbb{P}\{\eta > y\}.$$

Letting $n \to \infty$ and taking into account the duality between the single server system and the maximum of the corresponding random walk, we arrive at the following inequality:

$$\mathbb{P}\{D > x+y\} \leq \left(\begin{array}{c} s \\ k \end{array} \right) (\overline{F}(x))^{s-k} + \mathbb{P}\{\eta > y\}. \tag{33}$$

Since η has a finite exponential moment, we obtain the statement of the theorem. □

PROOF OF THE UPPER BOUND IN THEOREM 1.6. It follows from Theorem 7.1 that

$$\mathbb{P}\{D > x\} \leq \left(\begin{array}{c} s \\ k \end{array} \right) \overline{F}^{s-k}(x(1-\delta)) + const \cdot e^{-\beta\delta x}.$$

Due to the subexponentiality of B_r we obtain from the analogue of (2) for the maximum of a random walk (see, e.g., [10, Theorem 5.2]) that, as $x \to \infty$,

$$\overline{F}(x) \sim \frac{b}{(k+1)(a-h)-b} \overline{B}_r(x).$$

Taking h in (23) close to its minimal value, say $h = \frac{k}{k+1}\left(a - \frac{b}{k+1}\right) + \varepsilon$, $\varepsilon > 0$, we arrive at the following estimate:

$$\overline{F}(x) \sim \frac{b}{a - b/(k+1) - (k+1)\varepsilon} \overline{B}_r(x) = \frac{(k+1)\rho}{k+1-\rho-(k+1)^2\varepsilon/a} \overline{B}_r(x).$$

In addition, $\overline{B}_r(x(1-\delta)) \cdot e^{\beta\delta x} \to \infty$ as $x \to \infty$. All these facts and arbitrarity of choice of $\varepsilon > 0$ imply the desired bound. □

It what follows, for two families of events A_x and B_x of positive probabilities indexed by x, we write $A_x \sim B_x$ if $\mathbb{P}\{A_x \setminus B_x\} = o(\mathbb{P}\{A_x\})$ and $\mathbb{P}\{B_x \setminus A_x\} = o(\mathbb{P}\{A_x\})$ as $x \to \infty$. Note that $A_x \sim B_x$ implies $\mathbb{P}\{A_x\} \sim \mathbb{P}\{B_x\}$, but not vice versa.

We establish now the principle of $s-k$ big jumps in the case of intermediate regularly varying distributions. For simplicity, we do it again for $D/GI/s$ system with deterministic inter-arrival times. For this, we consider the representation of the stationary workload in the backward time (the so-called "Loynes scheme"). We again use the joint representation of s individual queues and of the s-server system given in the previous section and assume that all queues run for a long time, from time $-\infty$, and that U_i is the stationary waiting time of the customer that arrives at the ith queue at time 0. Then

$$U_i = \sup(0, \xi_{-1,i}, \xi_{-1,i} + \xi_{-2,i}, \ldots, \xi_{-1,i} + \ldots + \xi_{-n,i}, \ldots)$$

where $\xi_{j,i} = \sigma_{j,i} - \widehat{a}$, for $i = 1, \ldots, s$ and $j = -1, -2, \ldots$, and $\widehat{a} = (k+1)(a-h)$. Further, $\mathbf{W}_n$ are stationary vectors for $n \leq 0$, and

$$\mathbf{W}_{n+1} = R((W_{n1} + \sigma_n - a)^+, (W_{n2} - a)^+, \ldots, (W_{n,s} - a)^+),$$

for all $n < 0$. Here again $i_n = \min\{i : \ V_{ni} = D_n\}$ and $\sigma_n = \sigma_{n,i_n}$. Then, for any $x > 0$, by Lemma 6.1,

$$\{D_0 > x\} \subset \bigcup_J \{\min_{i \in J} U_i + \eta > x\}$$

where D_0 is the stationary waiting time in the $D/GI/s$ queue, i.e. the minimal coordinate of vector $\mathbf{W}_0$, J's are subsets of $\{1, 2, \ldots, s\}$ of cardinality $s-k$, and η is a random variable with light-tailed distribution. Then

$$\{D_0 > x\} = \bigcup_J \{D_0 > x, \min_{i \in J} U_i + \eta > x\}.$$

Assume that the residual distribution function B_r of service times is intermediate regularly varying (for that, it is sufficient for B to be intermediate varying). Then, clearly, each random variable U_i has an intermediate regularly varying distribution since $\mathbb{P}(U_i > x) \sim c\overline{B}_r(x)$. Since the random variables U_i are mutually independent, the distribution of $\min_{i\in J} U_i$ is also intermediate regularly varying, $\mathbb{P}(\min U_i > x) \sim c^{s-k} \left(\overline{B}_r(x)\right)^{s-k}$.

It is well-known, see e.g. [10, Ch 5], that

$$\{U_i > x\} \sim \bigcup_{n\geq 1} \{\sigma_{-n,i} > x + n\widehat{a}\}$$

and therefore, for any set $J \subset \{1, 2, \ldots, s\}$,

$$\{\min_{i\in J} U_i > x\} \sim \bigcap_{i\in J} \left(\bigcup_{n>0} \{\sigma_{-n,i} > x + n\widehat{a}\} \right).$$

We use the following property of intermediate regularly varying distributions (its proof is postponed until the end of the section; a similar result for equivalence of probabilities may be found in [2]):

LEMMA 7.1 *If X and Y are two random variables such that X has an intermediate regularly varying distribution and $\mathbb{P}\{|Y| > x\} = o(\mathbb{P}\{X > x\})$ as $x \to \infty$, then $\{X + Y > x\} \sim \{X > x\}$, for any joint distribution of X and Y.*

Applying Lemma 7.1 with $X = \min_{i\in J} U_i$ and $Y = \eta$, for all J of cardinality $s - k$, we get

$$\begin{aligned} \{D_0 > x\} &\sim \bigcup_J \{D_0 > x, \min_{i\in J} U_i > x\} \\ &\sim \bigcup_J \bigcap_{i\in J} \left(\bigcup_{n>0} \{D_0 > x, \sigma_{-n,i} > x + n\widehat{a}\} \right), \end{aligned}$$

since the upper and the lower bounds for $\mathbb{P}\{D_0 > x\}$ are of the same order, see [3, Theorem 7] or [9] for further arguments. Now represent any event on the right in the latter equation as a union of two events

$$\{D_0 > x, \sigma_{-n,i} > x + n\widehat{a}, i_{-n} = i\} \cup \{D_0 > x, \sigma_{-n,i} > x + n\widehat{a}, i_{-n} \neq i\}$$

where

$$\begin{aligned} \mathbb{P}\{D_0 > x, \sigma_{-n,i} > x + n\widehat{a}, i_{-n} \neq i\} &= \mathbb{P}\{D_0 > x, i_{-n} \neq i\}\mathbb{P}\{\sigma_{-n,i} > x + n\widehat{a}\} \\ &\leq \mathbb{P}\{D_0 > x\}\mathbb{P}\{\sigma_{-n,i} > x + n\widehat{a}\}. \end{aligned}$$

So, for any set J, the union of events

$$\bigcap_{i\in J} \left(\bigcup_{n>0} \{D_0 > x, \sigma_{-n,i} > x + n\widehat{a}, i \neq i_{-n}\} \right)$$

has probability $O(\mathbb{P}\{D_0 > x\}\overline{B}_r(x)) = o(\mathbb{P}\{D_0 > x\})$. Since there is only a finite number of sets J, we obtain the following result.

THEOREM 7.2 *Assume that $\rho \in (k, k + 1)$ and that the distribution of service times is intermediate regularly varying. As $x \to \infty$,*

$$\mathbb{P}\{D_0 > x\} \sim \mathbb{P}\Big\{D_0 > x, \bigcup_{0<n_1<n_2<\ldots<n_{s-k}} \bigcap_{j=1}^{s-k} \{\sigma_{-n_j} > x + n_j\widehat{a}\}\Big\}. \tag{34}$$

Remark. In the proof of Theorem 7.2, we followed the scheme introduced in [9], see also [3], [4], and [5] for similar constructions. Theorem 7.2 is not the final statement. We may go further and obtain the following result. Assume that B is a regularly varying distribution. Then, for some positive and finite constant C and as $x \to \infty$,

$$\mathbb{P}\{D_0 > x\} \sim C\overline{B}_r^{s-k}(x). \tag{35}$$

The result seems to be correct, but its proof would be very lengthy and would require a scrupulous calculation, so we decided not to proceed further in this direction.

We provide a hint for a plausible proof instead. First, one may consider an auxiliary deterministic model with $(n-s)$ very big service times $y_1, \ldots, y_{s-k}$ that occur at time instants $-n_1 > -n_2 > \ldots > -n_{s-k}$ and replace all other service times by their mean b. We also assume that, before the first jump, the workload vector is zero. For this model, we may find conditions on the y's for the minimal coordinate of the workload vector at time 0 to be not smaller than x. Then repeat the same for all the other times of jumps $-n_1 > -n_2 > \ldots > -n_{s-k}$. The union of these regions may be represented as a combination of unions and differences of a finite number of truncated half-spaces of dimension $s-k$. Summation of tail probabilities over each such set gives the probability of order $\overline{B}_r^{s-k}(x)$, thanks to the properties of regularly varying functions. So a finite combination of sums and differences of these probabilities gives a probability of the same order. It cannot be of a lower order, due to the lower bound.

Proof of Lemma 7.1. From Theorem 2.47 in [10], if X has an intermediate regularly varying distribution, then

$$\mathbb{P}\{X > x + h(x)\} \sim \mathbb{P}\{X > x\} \sim \mathbb{P}\{X > x - h(x)\}$$

as $x \to \infty$, for any function $h(x) \to \infty$ such that $h(x) = o(x)$. Hence, by the monotonicity arguments,

$$\{X > x + h(x)\} \sim \{X > x\} \sim \{X > x - h(x)\}.$$

Since the distribution of X is intermediate regularly varying and since $\mathbb{P}\{|Y| > x\} = o(\mathbb{P}\{X > x\})$ as $x \to \infty$, we have $\mathbb{P}\{|Y| > \varepsilon x\} = o(\mathbb{P}\{X > x\})$ as $x \to \infty$, for every $\varepsilon > 0$. Then there exists $h(x) = o(x)$ such that

$$\mathbb{P}\{|Y| > h(x)\} = o(\mathbb{P}\{X > x\}).$$

Therefore, as $x \to \infty$,

$$\begin{aligned} \{X > x\} &\sim \{X > x + h(x)\} \setminus \{Y < -h(x)\} \\ &= \{X > x + h(x), Y \geq -h(x)\} \subseteq \{X + Y > x\} \end{aligned}$$

and

$$\{X > x\} \sim \{X > x - h(x)\} \cup \{Y > h(x)\} \supseteq \{X + Y > x\},$$

which justifies the events equivalence, $\{X + Y > x\} \sim \{X > x\}$. □

8. Existence of moments: proof of Theorem 4. Since the tail distribution of $\min(\sigma_{r,1}, \ldots, \sigma_{r,s-k})$ is equal to $(\overline{B}_r(x))^{s-k}$, we obtain from Theorem 5.1

$$\mathbb{P}\{D > x\} \geq c_1 \mathbb{P}\{\min(\sigma_{r,1}, \ldots, \sigma_{r,s-k}) > c_2 x\}.$$

Since, for any non-negative random variable η,

$$\mathbb{E}\eta^\gamma = \gamma \int_0^\infty x^{\gamma-1} \mathbb{P}\{\eta > x\} dx,$$

we have

$$\mathbb{E}D^\gamma \geq \frac{c_1}{c_2} \mathbb{E}(\min(\sigma_{r,1}, \ldots, \sigma_{r,s-k}))^\gamma.$$

and the existence of the moment of order γ for the delay D implies with necessity (4).

Now assume (4). Consider $s-k$ independent copies $M_1, \ldots, M_{s-k}$ of the random variable M introduced in Theorem 7.1. Then the assertion of Theorem 7.1 can be rewritten in the following way:

$$\mathbb{P}\{D > x + y\} \leq \binom{s}{k} \mathbb{P}\{\min(M_1, \ldots, M_{s-k}) > x\} + const \cdot e^{-\beta y}.$$

Take $y = x$. Then $\mathbb{E}D^\gamma < \infty$ follows if we prove that

$$\mathbb{E}(\min(M_1, \ldots, M_{s-k}))^\gamma < \infty. \tag{36}$$

In order to do it, we explore the ladder height construction for the maximum M of a random walk $S_n = X_1 + \ldots + X_n$ where $X_j = \sigma_j - b - \varepsilon$. Since this random walk has a negative drift, the first ladder epoch and the first ladder height

$$\theta = \min(n \geq 1 : S_n > 0), \quad \widetilde{\chi} = S_\theta,$$

both are degenerate random variables;

$$p \equiv \mathbb{P}\{\theta < \infty\} = \mathbb{P}\{M > 0\} < 1.$$

Denote by χ a random variable with distribution

$$\mathbb{P}\{\chi \in B\} = \mathbb{P}\{\widetilde{\chi} \in B\}/p.$$

Let χ_j be independent copies of χ. If η is an independent counting random variable with distribution $\mathbb{P}\{\eta = j\} = (1-p)p^j$, $j = 0, 1, \ldots$, then M is equal in distribution to $\chi_1 + \ldots + \chi_\eta$.

Let $\chi_{i,j}$ be again independent copies of χ and η_j be independent copies of η. Then $\min(M_1, \ldots, M_{s-k})$ is equal in distribution to

$$\min\Bigl(\sum_{j=1}^{\eta_1} \chi_{1,j}, \ldots, \sum_{j=1}^{\eta_{s-k}} \chi_{s-k,j}\Bigr).$$

The latter minimum does not exceed

$$\sum_{j_1=1}^{\eta_1} \ldots \sum_{j_{s-k}=1}^{\eta_{s-k}} \min(\chi_{1,j_1}, \ldots, \chi_{s-k,j_{s-k}}).$$

Taking into account that for non-negative arguments

$$(x_1 + \ldots + x_N)^\gamma \le N^\gamma (x_1^\gamma + \ldots + x_N^\gamma),$$

we get the following estimate:

$$\min\Bigl(\sum_{j=1}^{\eta_1} \chi_{1,j}, \ldots, \sum_{j=1}^{\eta_{s-k}} \chi_{s-k,j}\Bigr)^\gamma \quad \le \quad (\eta_1 + \ldots + \eta_{s-k})^\gamma \sum_{j_1=1}^{\eta_1} \ldots \sum_{j_{s-k}=1}^{\eta_{s-k}} \min(\chi_{1,j_1}, \ldots, \chi_{s-k,j_{s-k}})^\gamma.$$

In particular, the mean of the term in the left side of the equality above is not larger than

$$\sum_{j_1=1}^{\infty} \ldots \sum_{j_{s-k}=1}^{\infty} (1-p)p^{j_1+\ldots+j_{s-k}} (j_1 + \ldots + j_{s-k})^\gamma j_1 \cdot \ldots \cdot j_{s-k} \mathbb{E} \min(\chi_{1,1}, \ldots, \chi_{s-k,1})^\gamma$$
$$= \mathbb{E}(\eta_1 + \ldots + \eta_{s-k})^\gamma \eta_1 \ldots \eta_{s-k} \mathbb{E} \min(\chi_{1,1}, \ldots, \chi_{s-k,1})^\gamma.$$

Since the η's have finite exponential moments, the first mean on the right is finite. Now we show finiteness of the second mean. First,

$$\mathbb{P}\{\chi > x\} = \int_{-\infty}^{0} \overline{B}(x-y)\mu(dy),$$

where the measure μ is defined by

$$\begin{aligned} \mu(dy) &= \sum_n \mathbb{P}\{S_n \in dy, S_k \le 0 \text{ for all } k \le n-1\} \\ &\le \sum_n \mathbb{P}\{S_n \in dy\}. \end{aligned}$$

Then, by the key renewal theorem,

$$c \equiv \sup_{y \le 0} \mu(y-1, y] < \infty,$$

which yields

$$\mathbb{P}\{\chi > x\} \le c \sum_{j=0}^{\infty} \overline{B}(x+j) \le c\overline{B}_r(x-1).$$

Therefore, due to condition (4),

$$\mathbb{E} \min(\chi_{1,1}, \ldots, \chi_{s-k,1})^\gamma < \infty,$$

which completes the proof.

Acknowledgments. This research was supported by EPSRC grant No. R58765/01 and RFBR grant No. 10-01-00161. The authors thank Bert Zwart for stimulating discussions, and Arcady Shemyakin and James Cruise for stylistic comments.

References

[1] S. Asmussen, *Applied Probability and Queues*, 2nd ed. Springer, New York, 2003.

[2] H. Albrecher, S. Asmussen and D. Kortschak, *Tail asymptotics for dependent subexponential differences*, Sib. Math. J. **53** (2012), to appear.

[3] F. Baccelli and S. Foss, *Moments and tails in monotone-separable stochastic networks*, Ann. Appl. Probab. **14** (2004), 612–650.

[4] S. Borst, M. Mandjes and A. P. Zwart, *Exact asymptotics for fluid queues fed by heavy-tailed On-Off flows*, Ann. Appl. Probab. **14** (2004), 903–957.

[5] S. Borst and B. Zwart, *Fluid queues with heavy-tailed $M/G/\infty$ input*, Math. Oper. Res. **30** (2005), 852–879.

[6] O. J. Boxma, S. G. Foss, J.-M. Lasgouttes and R. Nunez Queija, *Waiting time asymptotics in the single server queue with service in random order*, Queueing Systems **46** (2004), 35–73.

[7] O. J. Boxma, Q. Deng and A. P. Zwart, *Waiting-time asymptotics for the $M/G/2$ queue with heterogeneous servers*, Queueing Systems **40** (2002), 5–31.

[8] H. Cramér, *Collective risk theory*, Esselte, Stockholm, 1955.

[9] S. Foss and D. Korshunov, *Heavy tails in multi-server queues*, Queueing Systems **52** (2006), 31–48.

[10] S. Foss, F. Korshunov and S. Zachary, *An Introduction to Heavy-Tailed and Subexponential Disttributions*, Springer, New York, 2011.

[11] J. Kiefer and J. Wolfowitz, *On the theory of queues with many servers*, Tran. Amer. Math. Soc. **78** (1955), 1–18.

[12] J. Kiefer and J. Wolfowitz, *On the characteristics of the general queueing process with applications to random walk*, Ann. Math. Stat. **27** (1956), 147–161.

[13] D. V. Lindley, *The theory of queues with a single server*, Proc. Cambridge Philos. Soc. **8** (1952), 277–289.

[14] A. G. Pakes, *On the tails of waiting-time distribution*, J. Appl. Probab. **12** (1975), 555–564.

[15] A. Scheller-Wolf, *Further delay moment results for FIFO multiserver queues*, Queueing Systems **34** (2000), 387–400.

[16] A. Scheller-Wolf and K. Sigman, *Delay moments for FIFO $GI/GI/s$ queues*, Queueing Systems **25** (1997), 77–95.

[17] A. Scheller-Wolf and R. Vesilo, *Structural interpretation and derivation of necessary and sufficient conditions for delay moments in FIFO multiserver queues*, Queueing Systems **54** (2006), 221–232.

[18] A. Scheller-Wolf and R. Vesilo, *Sink or swim together: necessary and sufficient conditions for finite moments of workload components in FIFO multiserver queues*, Queueing Systems **67** (2011), 47–61.

[19] N. Veraverbeke, *Asymptotic behavior of Wiener-Hopf factors of a random walk*, Stochastic Process. Appl. **5** (1977), 27–37.

[20] W. Whitt, *The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution*, Queueing Systems **36** (2000), 71–87.